Informed Source Separation from compressed mixtures using spatial wiener filter and quantization noise estimation

Shuhua Zhang, Laurent Girin, Antoine Liutkus

▶ **To cite this version:**

Shuhua Zhang, Laurent Girin, Antoine Liutkus. Informed Source Separation from compressed mixtures using spatial wiener filter and quantization noise estimation. ICASSP 2013 - 38th IEEE International Conference on Acoustics, Speech and Signal Processing, May 2013, Vancouver, Canada. pp.61-65, 10.1109/ICASSP.2013.6637609 . hal-00940328

HAL Id: hal-00940328
https://hal.science/hal-00940328

Submitted on 31 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INFORMED SOURCE SEPARATION FROM COMPRESSED MIXTURES USING SPATIAL WIENER FILTER AND QUANTIZATION NOISE ESTIMATION

*Shuhua Zhang, Laurent Girin**

GIPSA-Lab, Grenoble-INP
Grenoble, France

Antoine Liutkus

Telecom ParisTech, Institut Mines-Telecom, CNRS LTCI
Paris, France

ABSTRACT

In a previous work, we proposed an Informed Source Separation system based on Wiener filtering for active listening of music from uncompressed (16-bit PCM) multichannel mix signals. In the present work, the system is improved to work with (MPEG-2 AAC) compressed mix signals: quantization noise is estimated from the AAC bitstream at the decoder and explicitly taken into account in the source separation process. Also a direct MDCT-to-STFT transform is used to optimize the computational efficiency of the process in the STFT domain from AAC-decoded MDCT coefficients.

Index Terms— Informed Source Separation, Wiener Filter, Denoising, NTF, AAC

1. INTRODUCTION

Active listening of music is a topic of both artistic and technological interest that consists in modifying the elements and structure of the music signal during the listening experience. It involves online advanced remixing processes such as generalized karaoke (the ability to mute any musical element, hence not limited to voice), respatialization, or application of instrument-specific audio effects (e.g., adding distortion to an acoustic guitar). To enable active listening from compliant 2-channel stereo mix signals, recent technologies have been proposed. In Spatial Audio Object Coding (SAOC) [1], standardized by MPEG, the stereo mix signal is enriched with spatial coding parameters at the coder stage, that enable to decode the multitrack source signals from the downmix signal at the decoder stage. Recently, this idea was revisited from the *source separation* [2] point of view (see [3] and references therein). In particular, a Wiener-based Informed Source Separation (ISS) system was proposed in [4], where the side-information is composed of power spectral density (PSD) models of the source signals plus the mixing matrix, and separating Wiener filters are used at the decoder.

The new contributions of this paper with respect to previous works are threefold. First and more importantly, although ISS (and SAOC) purpose is to reduce the overall bitrate compared to multitrack coding (in addition to 2-channel stereo compliance with usual formats), the effects of *compressing the mix signal* on source signals decoding/separation is not being taken into account in those state-of-the-art methods, leading to suboptimal performance. In the present paper, we present an improved version of the Wiener-based ISS system, where the mix signal is compressed, for instance using MPEG-2 Advanced Audio Coding (AAC [5]), and the resulting quantization noise is explicitly accounted for in the separation process. This enables to significantly reduce the distortion in the separated source signals. Moreover, we propose to estimate this quantization noise at the decoder, so that no additional side-information needs to be transmitted. Second, we use here the *spatial* Wiener filtering scheme, as recently introduced in the ISS context in [6], which permits to fully benefit from the spatial dispersion of the sources within the mix to improve separation. And finally, the recently developed direct MDCT-to-STFT conversion technique of [7] is used to enable fast and efficient STFT-domain processing directly from AAC-decoded MDCT coefficients.

This paper is organized as follows. First, the proposed separation technique and system are described in Section 2. Then explicit handling of the AAC quantization noise is presented in Section 3. The improved system is shown in Section 4 to provide significantly better quality compared to when the quantization noise is not taken into account, for identical side-information rate.

*This work is supported by the French National Research Agency (ANR) as part of the DReaM project — ANR CONTINT 2009-006.

2. SYSTEM OVERVIEW

2.1. Spatial Wiener Filtering

The core source separation module of the proposed ISS system is based on the so-called spatial Wiener filter, or multichannel MMSE linear estimator [6, 2]. Let us assume that we have a J-source signals vector $\boldsymbol{s}[n] = \left[s_1[n]\ s_2[n]\ \cdots\ s_J[n]\right]^\top$ (n for time sample index, $^\top$ for transposition), which is downmixed ($I < J$) into a I-channel vector $\boldsymbol{x}[n] = \left[x_1[n]\ x_2[n]\ \cdots\ x_I[n]\right]^\top$ using linear time-invariant (LTI) filters, i.e. underdetermined convolutive mixing. It is classical to process the separation problem, i.e. recovering of $\boldsymbol{s}$ from $\boldsymbol{x}$, in the frequency domain, where the mixture may be approximated as instantaneous at each frequency f [8], using Short-Time Fourier Transform (STFT) (t denotes the frame index):

$$\boldsymbol{X}(f,t) = \mathbf{A}(f)\boldsymbol{S}(f,t) + \boldsymbol{N}(f,t), \tag{1}$$

where $\boldsymbol{X}(f,t) = \text{STFT}\{\boldsymbol{x}[n]\}$, $\boldsymbol{S}(f,t) = \text{STFT}\{\boldsymbol{s}[n]\}$, $\mathbf{A}(f)$ is the $I \times J$ frequency-dependent downmix matrix, and $\boldsymbol{N}(f,t)$ is the noise. Here, the sources are assumed to be independent zero-mean complex Gaussian processes, i.e., $\boldsymbol{S}(f,t) \sim \mathcal{N}(0, \mathbf{R}_{\boldsymbol{SS}}(f,t))$, where $\mathbf{R}_{\boldsymbol{SS}}(f,t)$, the $J \times J$ covariance matrix of the sources, is diagonal and the j-th diagonal term is the PSD of source j at TF bin (f,t). The noise is also assumed to be Gaussian: $\boldsymbol{N}(f,t) \sim \mathcal{N}(0, \mathbf{R}_{\boldsymbol{NN}}(f,t))$. Let us denote $\mathbf{R}_{\boldsymbol{SX}}(f,t) \equiv \mathbf{R}_{\boldsymbol{SS}}(f,t)\mathbf{A}(f)^\top$. Under the above parameters and given the mixture $\boldsymbol{X}(f,t)$, we have the MMSE estimator $\widehat{\boldsymbol{S}}(f,t)$ of the source signals vector:

$$\widehat{\boldsymbol{S}}(f,t) = \mathbf{W}(f,t)\boldsymbol{X}(f,t), \tag{2}$$

with

$$\mathbf{W}(f,t) \equiv \mathbf{R}_{\boldsymbol{SX}}(f,t)[\mathbf{A}(f)\mathbf{R}_{\boldsymbol{SX}}(f,t) + \mathbf{R}_{\boldsymbol{NN}}(f,t)]^{-1}. \tag{3}$$

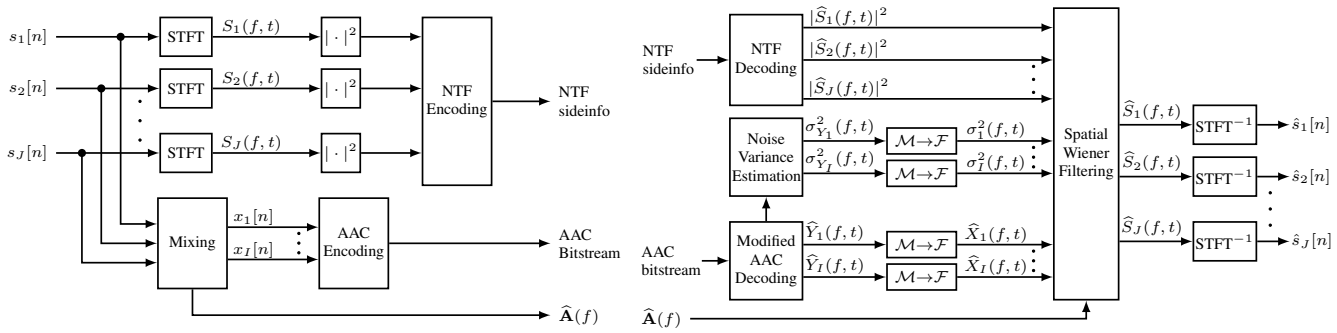

Fig. 1. Functional representation of the spatial Wiener filter based ISS system: encoder (left) and decoder (right). Here $\mathcal{M} \to \mathcal{F}$ is for MDCT-to-STFT conversion, see Section 3.3.

2.2. ISS encoder and decoder

The ISS encoder based on the spatial Wiener filter is schematized in Fig. 1-left. The inputs to the encoder are the source signals $s_j[n]$ which are both converted to STFT coefficients $S_j(f,t)$ and (convolutively) downmixed into $\boldsymbol{x}[n]$. $\mathbf{A}(f)$ is quantized to be transmitted to the decoder as part of the side-information. The AAC compression stage for $\boldsymbol{x}[n]$ is addressed in Section 3. Assuming the sources are locally stationary Gaussian processes, the maximum likelihood estimator of PSD is spectrogram $|S_j(f,t)|^2$ [9], so that we redefine $\mathbf{R}_{\boldsymbol{SS}}(f,t) = \text{diag}\,\{|S_1(f,t)|^2, |S_2(f,t)|^2, \ldots, |S_J(f,t)|^2\}$. This covariance matrix also has to be transmitted to the decoder for Wiener filter construction. However, spectrograms generally occupy a very large amount of side-information, so they must be substantially compressed. This is done using Non-negative Tensor Factorization (NTF) modeling of the spectrograms, arranged as 3D tensors (frequency $\times$ time $\times$ source), as in [4, 6], and subsequent NTF parameters quantization.

The ISS decoder side is shown in Fig. 1-right. The NTF models of spectrograms and mixing matrix are decoded from the side-info to provide $\widehat{\mathbf{R}}_{\boldsymbol{SS}}(f,t)$ and $\widehat{\mathbf{A}}(f)$ from which the Wiener filter (3) is built. Filtered source coefficients $\widehat{S}_1(f,t), \widehat{S}_2(f,t), \ldots, \widehat{S}_J(f,t)$ are then obtained from mix coefficients $\widehat{X}_1(f,t), \ldots, \widehat{X}_I(f,t)$, and synthesized to time-domain signals $\hat{s}_1[n], \hat{s}_2[n], \ldots, \hat{s}_J[n]$ using inverse STFT with overlap-add. Other processes are described in the next sections.

3. NOISE VARIANCE ESTIMATION FROM AAC BITSTREAM

In our previous works [4, 6], 16-bit PCM mix signal $\boldsymbol{x}$ was processed at the coder and decoder. Hence it was considered as unquantized and the noise $\boldsymbol{N}(f,t)$ in (1) was assumed to be null[1]. In the present work, lossy MPEG-2 AAC compression [5] is applied on the mix signal, introducing compression noise (mainly due to quantization) at each TF bin. Therefore, we propose to specifically model the degradation of the mixture due to compression as the additive noise $\boldsymbol{N}(f,t)$ in (1), independent from the sources. Doing so, we allow for the separation procedure (3) to take it into account to better recover the sources. This noise $\boldsymbol{N}(f,t)$ may be estimated simply at the coder as the difference between original and compressed mixture, and its PSD may be encoded and transmitted to the decoder using the NTF model. However, this would increase the side-information bitrate. Instead, we show below how the noise variance can be estimated directly from AAC bitstream.

3.1. Nonlinear Quantization Noise

In AAC, compression is based on the quantization of Modified Discrete Cosine Transform (MDCT) [10] coefficients. A coefficient $Y(k)$ on the k-th frequency bin (frame index is omitted here for clarity) within subband b is quantized to an integer value by scaling, $\frac{3}{4}$-power compression, and rounding:

$$\bar{Y}(k) = \text{sgn}\,(Y(k)) \left\lfloor (|Y(k)|/2^{\frac{1}{4}s(b)})^{\frac{3}{4}} + C \right\rfloor, \qquad (4)$$

where $C \equiv 1 - 2^{-3/4}$ and $s(b)$ is the scalefactor for subband b. Now suppose $\bar{Y}(k) = q$ is positive. Then by (4), any MDCT coefficient $Y(k) \in 2^{\frac{1}{4}s(b)}[T_q, T_{q+1})$ will be quantized to q, where $T_q \equiv (q - C)^{4/3}$ is the lower boundary of the q-th quantization interval. A similar result can be derived for $q \leq 0$:

$$Y(k) \in 2^{\frac{1}{4}s(b)} \times \begin{cases} (-T_{|q|+1}, -T_{|q|}], & q < 0, \\ (-T_1, \quad T_1), & q = 0, \\ [\ T_q, \quad T_{q+1}), & q > 0. \end{cases} \qquad (5)$$

Due to the $\frac{3}{4}$-power compression, this quantization is nonlinear and the intervals have increasing lengths as $|q|$ increases. Furthermore, assuming MDCT coefficients are uniformly distributed on each interval, the noise variance is

$$\sigma^2(k) = \begin{cases} \frac{1}{12}(T_{|q|+1} - T_{|q|})^2 2^{\frac{1}{2}s(b)}, & q \neq 0, \\ \frac{1}{12}(2T_1)^2 2^{\frac{1}{2}s(b)}, & q = 0. \end{cases} \qquad (6)$$

Thus, after retrieving quantization index q and scalefactor $s(b)$ from an AAC bitstream, (6) can be used at the decoder to estimate the variance of the quantization noise for a single channel[2].

3.2. Stereo Coding Noise

In AAC, each channel is either encoded independently or paired with another channel for stereo coding. In the latter case, both redundancy and irrelevancy between the paired channels (left and right)

[1] Actually, $\boldsymbol{N}(f,t)$ in (1) can be used to model the mixing model error (resulting from STFT of convolution into product) when mixing filters are long, but this is very different and independent from compression noise.

[2] Another tool, called Temporal Noise Shaping (TNS [11]), changes noise distribution. We ignore this tool here for simplicity. Experiments show this neglect does not noticeably impair sound quality of separated sources.

are exploited to save bitrate using either the Mid/Side (M/S) [12] or Intensity Stereo (IS) [13] coding tools. We shall consider these two tools separately in the following.

M/S coding is noiseless but changes the noise distribution. Suppose that M/S is applied at bin k. The coefficients to be quantized are the middle and side of the MDCT coefficients $Y_L(k)$ and $Y_R(k)$ from left and right channels: $Y_0(k) = \frac{1}{2}(Y_L(k) + Y_R(k))$ and $Y_1(k) = \frac{1}{2}(Y_L(k) - Y_R(k))$. The middle $Y_0(k)$ and side $Y_1(k)$ are quantized by (4) and their quantization noises follow (6). Assuming the noises are independent, then noise variances of the left and right channels are

$$\sigma_L^2(k) = \sigma_R^2(k) = \sigma_0^2(k) + \sigma_1^2(k), \tag{7}$$

in which $\sigma_0^2(k)$ and $\sigma_1^2(k)$ are the noise variances determined by (6) for $Y_0(k)$ and $Y_1(k)$, respectively.

Conversely, IS coding is noisy, introducing noise while changing noise distribution. Let $Y_S(k) = Y_L(k) + Y_R(k)$ be the sum channel and denote by $I_{L/S}(b)$ and $I_{L/R}(b)$ the ratios of intensities (square root of total energy) between the left and the sum channels, and between the left and the right channels, respectively, for subband b. In IS encoding, a pair of new left and right channels are built from the sum channel and the intensity ratios:

$$Y_l(k) = Y_S(k) I_{L/S}(b), \quad Y_r(k) = Y_l(k)/I_{L/R}(b). \tag{8}$$

Suppose within subband b, $Y_L(k)$ and $Y_R(k)$ are sufficiently correlated such that $Y_L(k) \approx Y_l(k)$ and $Y_R(k) \approx Y_r(k)$ (roughly the point of applying IS). Then, only the new left $Y_l(k)$ for each bin k and the intensity ratio $I_{L/R}(b)$ for the whole subband b are needed for the decoder, substantially saving bitrate. The new left $Y_l(k)$ will be quantized by (4) with noise variance $\sigma_l^2(k)$ determined by (6). The intensity ratio $I_{L/R}(b)$ will also be quantized to the so called IS position $p(b)$ in the logarithmic domain:

$$p(b) = \left\lfloor 4\log_2(I_{L/R}) + \tfrac{1}{2} \right\rfloor = 4\log_2(I_{L/R}) + \delta p(b), \tag{9}$$

in which $\delta p(b)$ is the rounding error assumed to be uniformly distributed on $(-0.5, 0.5]$. Thus the L/R noise variances can be estimated as

$$\begin{cases} \sigma_L^2(k) \approx \sigma_l^2(k), \\ \sigma_R^2(k) \approx 2^{-p(b)/2}\sigma^2(k) + B 2^{-p(b)/2} \widehat{Y}_l^2(k), \end{cases} \tag{10}$$

in which $B = 0.0037$ is the variance of $2^{\delta p(b)}$, and $\widehat{Y}_l(k)$ is $Y_l(k)$ after quantization and dequantization.

3.3. MDCT-to-STFT Conversion

Decoding of AAC bitstream is done in the MDCT domain. But the spatial Wiener filter building and filtering process is to be carried out in the STFT domain, where time-domain convolutive mixing can be approximated as instantaneous mixing [14]. Therefore, we propose to use the direct MDCT-to-STFT conversion technique of [7] to convert both the estimated quantization noise variances and the mix signal coefficients. Accordingly, the AAC decoder is modified such that it outputs the quantized MDCT coefficients $\hat{Y}_{L/R}(f,t)$ and other AAC parameters necessary for estimation and conversion (quantization indexes, scalefactors, stereo parameters, window shapes and sequences), and skip the MDCT-to-time signal conversion. This saves computational complexity by avoiding full AAC decoding and STFT analysis.

In [7], STFT coefficient $X(f,t)$ for frame t is obtained from MDCT coefficients of the previous, current and next frames using a linear FIR filtering form:

$$\begin{aligned} X(f,t) = \phi(f) \sum_{k=0}^{M-1} \Big\{ & (-1)^k \big[h_0(f-k-1) + h_0(f+k)\big] Y(k,t) \\ & + \big[h_{-1}(f-k-1) + h_{-1}(f+k)\big] Y(k,t-1) \\ & + \big[h_{+1}(f-k-1) + h_{+1}(f+k)\big] Y(k,t+1) \Big\} \end{aligned} \tag{11}$$

for $f = 0, 1, \ldots, M$, where M is the MDCT coefficients vector size, equal to half the STFT framesize, $\phi(f) = \exp[-\frac{\pi}{2M}(1+M)k]$ is a pure phase function, and $h_0(f)$, $h_{-1}(f)$, and $h_{+1}(f)$ are filter taps depending on the MDCT and STFT windows. Assuming that the quantization noises are independent between different frames, frequencies and channels, the noise variance in the STFT domain is

$$\begin{aligned} \sigma^2(f,t) = \sum_{k=0}^{M-1} \Big\{ & \big|h_0(f-k-1) + h_0(f+k)\big|^2 \sigma_Y^2(k,t) \\ & + \big|h_{-1}(f-k-1) + h_{-1}(f+k)\big|^2 \sigma_Y^2(k,t-1) \\ & + \big|h_{+1}(f-k-1) + h_{+1}(f+k)\big|^2 \sigma_Y^2(k,t+1) \Big\}. \end{aligned} \tag{12}$$

It is also shown in [7] that $h_0(f)$, $h_{-1}(f)$, and $h_{+1}(f)$ all rapidly tend to 0 as $|f|$ increases. For example, with $M = 1024$, the KBD window [15] for the MDCT and the Hanning window for the STFT, keeping a total of only 20 taps in (11) leads to a conversion SNR higher than 60 dB. Therefore, we use a similar setup in the experiments.

4. EXPERIMENTS

4.1. Implementation and Experiment Settings

The proposed ISS system has been implemented in Matlab. The AAC encoder, however, is a binary one from Nero[3] known for its high audio coding quality. The AAC decoder is the open source[4] FAAD2 in C, which is modified as mentioned above.

A set of 14 musical excerpts (10s long, sampled at 44.1 kHz, 5 to 10 source signals) from the Quaero database[5] is used. For each excerpt, sources are downmixed to 2-channel stereo using Head Related Impulse Responses (HRIR) filters collected from the CIPIC database [16] (with order 200). The corresponding mixing matrix $\mathbf{A}(f)$ is transmitted to the decoder as single-precision floating-point numbers. Since $\mathbf{A}(f)$ does not depend on the length of an excerpt, the corresponding bitrate is assumed to be negligible for a complete song, and is not accounted for here.

For each source and each frame, a STFT spectrum may come from one long window size of 2048 or eight short windows size of 256. In the latter case, STFT coefficients of the short windows are intertwined from low to high frequency. To save bitrate, a cutoff frequency is set to 16 kHz. The resulting tensor of spectrograms are compressed by NTF whose parameters are linearly quantized (usually 8 bits) in the logarithmic domain [17]. NTF compression ratio is mainly controlled by the number of NTF components [4], and values of $5, 10, 20, 30, 50$, are tested here. The quantized NTF parameters are entropy coded before transmitting to the decoder, using here the data compression utility of the '`save`' Matlab function with option

[3] http://www.nero.com/enu/technologies-aac-codec.html
[4] http://www.audiocoding.com/faad2.html
[5] http://www.quaero.org

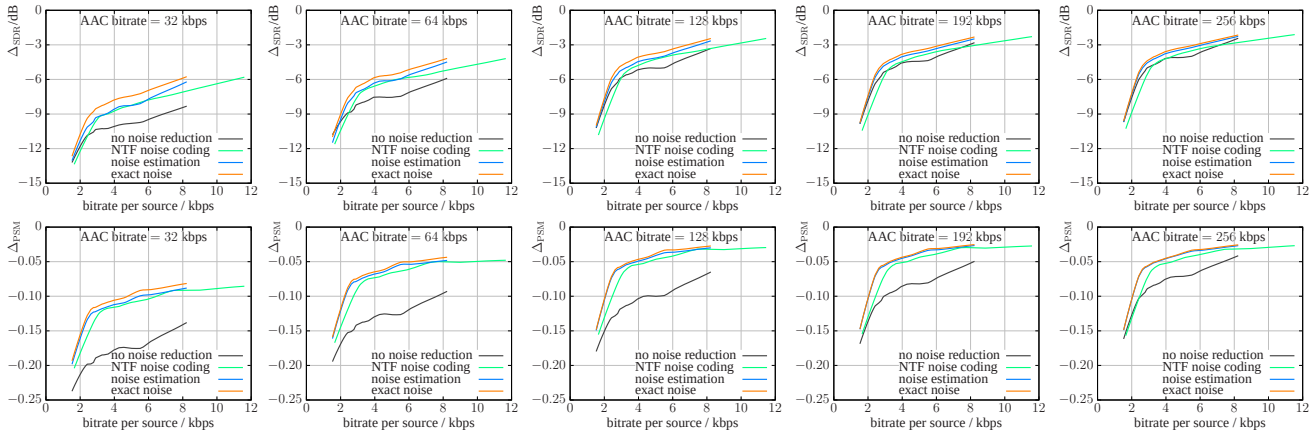

Fig. 2. Rate-SDR (top) and Rate-PSM (bottom) curves for the proposed Wiener-based ISS system.

'−V7' . Resulting NTF side-info bitrate per source ranges from 0.3 to 10 kbps.

The AAC bitrates that were tested for encoding the mix signal are $32, 96, 128, 192, 256$ kbps, covering the most common bitrates in AAC encoding for stereo (low complexity profile).

4.2. Results and Discussion

We used two metrics to measure separation performance: Perceptual Similarity Measure (PSM) of PEMO-Q [18], and Signal to Distortion Ratio (SDR) computed using the BSS Eval toolbox [19]. PSM provides an assessment of perceived quality of separated sources against the original sources, ranging from 0 (worst) to 1 (best). Since the performance heavily depends on excerpts, and to compare performance across all the test excerpts, each excerpt is finally measured by relative scores: $\Delta_{\mathrm{PSM}} \equiv$ PSM(oracle method) $-$ PSM(test method); $\Delta_{\mathrm{SDR}} \equiv$ SDR(oracle method) $-$ SDR(test method). The so called oracle method gives the highest possible performance using the spatial Wiener filter: Original spectrograms (instead of NTF compressed ones) are used to construct the separation filter and original mixtures (instead of AAC compressed ones) are input to the filter.

To study the effectiveness of quantization noise variance estimation and noise reduction in the ISS process, we run four configurations (all with NTF-compressed spectrograms and AAC compressed mixtures): (a) no noise reduction; (b) NTF coding of noise spectrograms + noise reduction; (c) noise variance estimation from AAC bitstream + noise reduction; (d) exact noise variance + noise reduction. The resulting Δ_{SDR} and Δ_{PSM} scores for all the 14 excerpts are smoothed to a single rate-quality curve by the LOWESS method [20] and then plotted against side-info rate in Fig. 2, for the five tested AAC bitrates. Each curve is then understood as a smoothing of 5×14 points (#NTF compression settings $\times$ #excerpts). Across the 14 excerpts, the mean PSM score of the oracle method is 0.882 (min 0.837, max 0.932), the mean SDR of the oracle method is 11.12 dB (min 7.10 dB, max 17.94 dB).

It can be seen from Fig. 2 that under all the tested AAC and side-info bitrates, configuration (c) performs better than configuration (a): increase of SDR upto 2.1 dB and increase of PSM upto 0.08,which represents substantial sound quality improvement, also confirmed by informal listening test. Quality improvement due to noise reduction generally decreases as AAC bitrate increases. This is because at higher bitrates, compression noise becomes smaller. Nevertheless, upto 128 kbps, the improvement in term of PSM is still significant (0.07). We also observed with noise reduction (configuration (c)), using bitrates higher than 128 kbps brings very limited quality gain. Thus we consider 128 kbps a sweet spot bitrate for AAC coding of the mixture under configuration (c).

Configuration (b) also improves quality over configuration (a), however, at the cost higher sideinfo rate, up to 4 kbps per source for the same improvement of configuration (c). Configuration (d) gives slightly better PSM scores (maximally 0.008 higher) and SDRs (maximally 0.9 dB higher) than configuration (c). The differences are generally imperceptible and shrink to 0 as AAC bitrate increases. This indicates that the estimated noise is accurate enough for the purpose of noise reduction. Obviously, configuration (d) is not realistic since exact noise is not available at the decoder in a practical system. Therefore, configuration (c) is always preferred.

This ISS system is efficient. Implemented mainly in Matlab (excluding AAC encoding and decoding), the system runs $\frac{1}{3}$X to $\frac{1}{2}$X real-time at the encoder side and 6X to 10X real-time at the decoder side (CPU 3.0 GHz). If implemented in C or other compiled languages, the system can be faster.

5. CONCLUSION

In this paper, we have proposed an informed source separation system which permits to handle mixtures compressed with standard audio coders. This was done through spatial Wiener filtering with compression noise reduction. Moreover, we showed that for the noise reduction, all the needed parameters, i.e., noise variance at each TF bin, can be estimated directly from the compressed mixture bitstream, without transmission of additional side-info. A large evaluation demonstrated that by taking in account of compression noise, the proposed system significantly increases perceptual quality of separated audio sources, at no cost of side-info bitrate, and the estimated noise variances provided almost identical separation quality as the exact noise variance. In the further, we shall drop the assumption that the compression noise is independent from the sources and consider not just compression noise but modeling errors: for example, long convolutive mixing (up to seconds) and nonlinear/nonstationary mixing, using the recently proposed coding-based informed source separation technique.

6. REFERENCES

[1] J. Engdegård, C. Falch, O. Hellmuth, J. Herre, J. Hilpert, A. Hölzer, J. Koppens, H. Mundt, H. Oh, H. Purnhagen, B. Resch, L. Terentiev, M. Valero, and L. Villemoes, "MPEG Spatial Audio Object Coding—the ISO/MPEG standard for efficient coding of interactive audio scenes," in *129th Audio Engineering Society Convention*, San Francisco, CA, 2010.

[2] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation - Independent Component Analysis and Applications*, Academic Press, 2010.

[3] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed source separation: a comparative study," in *European Signal Proc. Conf.*, Bucharest, Romania, 2012.

[4] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[5] ISO/IEC JTC1/SC29/WG11 MPEG, "Coding of moving pictures and audio, part 7: Advanced Audio Coding," Tech. Rep. ISO/IEC 13818-7, 2005.

[6] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *European Signal Proc. Conf.*, Bucharest, Romania, 2012.

[7] S. Zhang and L. Girin, "Fast and accurate direct MDCT to DFT conversion with arbitrary window functions," *IEEE Transactions on Audio, Speech, and Language Processing*, pending publication, available on ieeexplore.ieee.org.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[9] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, pp. 3155 –3167, July 2011.

[10] H. Malvar, *Signal processing with lapped transforms*, Artech House, Norwood, USA, 1992.

[11] J. Herre and J.D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (tns)," in *101st Audio Engineering Society Convention*, 1996.

[12] J.D. Johnston, "Perceptual transform coding of wideband stereo signals," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1989.

[13] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *96th Audio Engineering Society Convention*, 1994.

[14] E. Stein and R. Shakarchi, *Fourier analysis: an introduction*, Princeton University Press, Princeton, New Jersey, USA, 2003.

[15] M. Bosi and R. Goldberg, *Introduction to digital audio coding and standards*, Kluwer Academic Publishers, Norwell, USA, 2003.

[16] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2001.

[17] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2011.

[18] R. Huber and B. Kollmeier, "PEMO-Q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, july 2006.

[20] W. Cleveland and S. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, Sept. 1988.